A picture containing text

Description automatically generated

Explaining Trumps dominance with use the of socio-economic Data

Athens University of Economics and Business

Master of Science in Business Analytics

Statistics for BA II

Christos Vlassis

f2822204

Table of Contents

[1.Abstract 3](#_Toc127358483)

[2.Introduction 3](#_Toc127358484)

[3.Data exploration 4](#_Toc127358485)

[4.Modelling & Analysis 10](#_Toc127358486)

[4.1 Methodology & modelling comparison 10](#_Toc127358487)

[4.1.1 First Model 10](#_Toc127358488)

[4.1.2 Second Model 10](#_Toc127358489)

[4.1.3 Third Model 10](#_Toc127358490)

[4.1.4 Fourth model 10](#_Toc127358491)

[4.1.5 Fifth model 10](#_Toc127358492)

[4.1.6 Sixth model 11](#_Toc127358493)

[4.2 The Main Model 11](#_Toc127358494)

[4.2.1 The Model 11](#_Toc127358495)

[4.2.2 Analysis and coefficient interpretation 12](#_Toc127358496)

[4.3 Assumption Control 12](#_Toc127358497)

[Conclusions 13](#_Toc127358498)

[Appendix 14](#_Toc127358499)

# 1.Abstract

In our attempt to explain the dominance of Trump in the American elections and the characteristics of the counties that contribute to his ascendance in power we created six models.

We found, in our main model, six statistically significant variables (4.2.2) including the states as categorical variable. Also, we described their relationship with the dependent variable in (3.) to have a better understanding of our significant variables.

# 2.Introduction

Generalized linear Models (GLM) was introduced from John Nelder and Rober Wedderburn in 1972. Under the GLM other models can be found such as the linear regression, Logistic regression and the poison regression. In this paper the Logistic Regression will be used to analyse the effect of the demographic data of the United States of America to whether Trump got more than 50% of votes at each county.

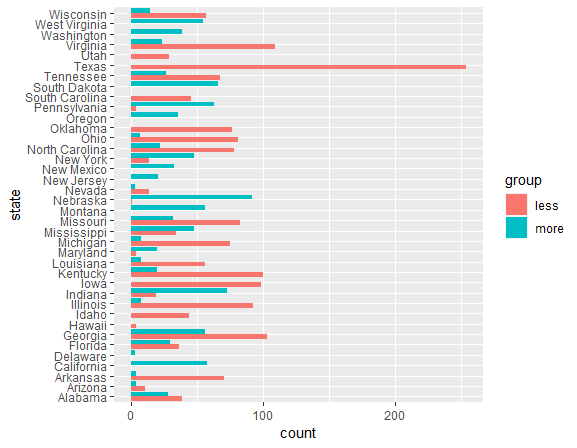
The following paper will be broken into two main parts. In the first part (Data exploration), we will investigate the most important variables and present the descriptive statistics of those variables. Also, we will make use of pairwise comparisons to understand their relationship with the response variable. Finally, we will explore, in brief, in which states we found Trump as the dominant candidate.

In the second part (Modelling & Analysis), we will investigate our response variable with the use of the logistic regression and we will create several models in our attempt to explain it as best we can.

# 3.Data exploration

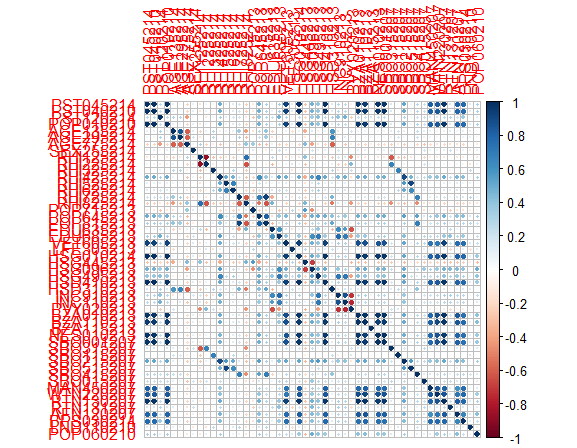
The data were given to us by Mr. Karlis and contain demographics for each county of the United States of America. There are 51 attributes in our disposal with 2711 observations. The purpose of this paper is to create a model only for inference so no training set and test set will be used. To have a better understanding of the data the following Graph counts for each state the counties with 50% more votes over Trump for the Republicans. With red we see the counties in which Trump was not dominant and with blue we count those that Trump was dominant. What we can derive from this Graph is that Trump dominated states such as Nebraska, Indiana, South Dakota, and Pennsylvania. On the other hand, in Virginia, Texas Arkansas Trump was not very popular and had a very small number of votes, less than 50%.

**Graph 1: Trumps dominance in state level**



Continuing with the descriptive analysis the independent variables are highly correlated. This will, later on, create issues regarding the multicollinearity assumption. With that being said it must be taken into consideration that many variables have the same information so, we will have to remove some of them to reduce the VIF scores.

**Graph 2: Correlation between variables**



# *The variable names has been removed. The graph demonstrates only the fact that there is high linear correlation between the independent variables. The full graph can be found in the appendix.*

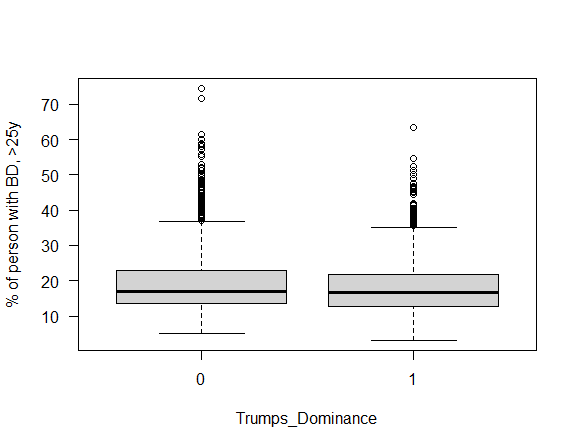
From the modelling experience that takes place in the second part of this paper we have derived with some variables that are statistically significant for most of the models that were tested and created. We will plot and describe these variables since they are the most important of all to understand their relationship and the way that they affect the response variable. Also, since showcasing all the independent variables is unfeasible, we will plot the most important that derive from the models. Finally, for each Graph we will use hypothesis testing to test if there is a significant difference between the response variable and the continuous variables.

**Table 1: Most Significant Variables**

|  |  |
| --- | --- |
| **Variable:** | **Description:** |
| AGE295214 | Persons under 18 years, % for 2014 |
| RHI325214 | American Indian and Alaska Native alone, %, 2014 |
| EDU685213 | Bachelor's degree or higher, % of persons age 25+ for 2009-2013 |
| PVY020213 | Persons below poverty level, %, 2009-2013 |

Because our response variable is a factor and the rest of the variables are continuous, the best way to describe the relationship between them is with the use of box plots or bar plots.

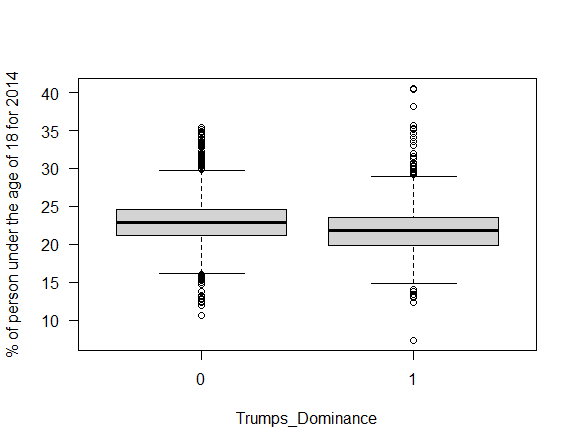
**Graph 3: Trumps dominance vs persons with Bachelors Degree, Age>25**



As can be seen from Graph 3 the median values for the percentage of persons with Bachelor’s degree and older than 25, for those county in which Trump was dominant have smaller median value, 16.6 unlike the counties in which Trump was not dominant, the median value was 17.05. This fact shows us that there is a small difference in the educational level of those counties that voted mostly for Trump. Finally, with a p-values of 0.81 we do not reject the null hypothesis for the Kruskal-Wallis test meaning that there is no statistically significant difference between the medians of the two groups

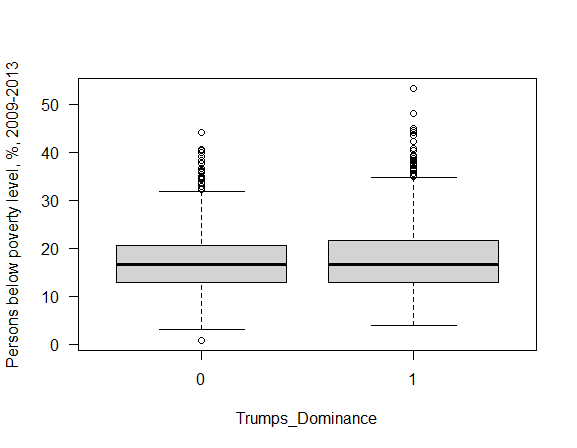
Continuous with our descriptive statistics the following Graphs shows the relationship between Trumps dominance and the percentage of persons under the age of 18 for 2014.

**Graph 4: Trumps dominance vs % of person under the age of 18 for 2014**



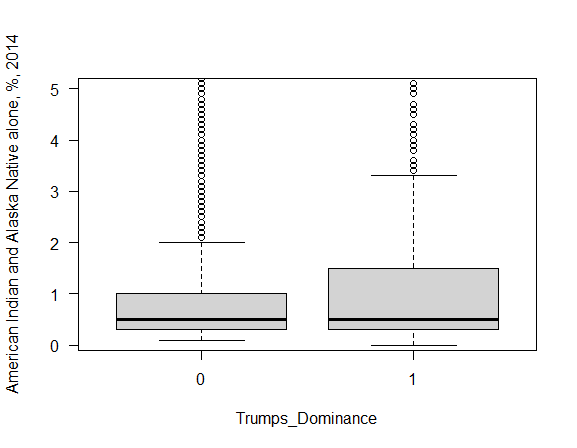
The median values for those counties that voted mostly for Trump have a smaller median (21.8). The counties that, on the most part, did not voted for trump had a median value of 22.9. Again, we can find a small difference indicating that more person under the age of 18 affect the negative Trumps dominance. Finally, with a p-values of 0.00003038 we reject the null hypothesis for the Kruskal-Wally’s test meaning that there is statistically significant difference between the medians of the two groups.

**Graph 5: Trumps dominance vs persons bellow povert level as %**



Graph 5 shows the persons bellow poverty level as % against the Trumps dominance in the counties. We can see that there is no big difference between the medians of the categories. Finally, with a p-values of 0.3784 we do not reject the null hypothesis for the Kruskal-Wallis test meaning that there is no statistically significant difference between the medians of the two groups.

**Graph 6: Trumps dominance vs % of American Indian & Alaska native for 2014**



In the Graph 6 we can see that there is no big difference in the medians of the American Indians and Alaska Natives, as percentage for 2014 between the counties in which Trump dominated. Finally, with a p-values of 0.0000001692 we reject the null hypothesis for the Kruskal-Wally’s test meaning that there is statistically significant difference between the medians of the two groups.

Concluding our descriptive analysis for this model, we found small differences in the medians of the most important variables of our model. But, we have to mention that the large number of outliers may give us misleading results. For further analysis those outliers should be removed, if possible, in order to have better results in this statistical research.

# 4.Modelling & Analysis

## 4.1 Methodology & modelling comparison

The main aim of this paper is the creation of a logistic model for inference. More specifically we want a model that can describe the characteristics that make Trump have more than 50% of votes in the USA counties.

In total, six models were created and tested all having some key differences. In some of them Lasso was used as a screening technique and later the BIC criterion for variable selection. To compare the six, in total, models we used three Pseudo R squares, ‘CoxSnell’, ‘Nagelkerke’, ‘McFaddenAdj’ and we also made use of ‘Hosmer-Lemeshow Goodness-of-Fit Test’ in order to access the goodness of fit of our models.

### 4.1.1 First Model

In the first model (Table 2, in the appendix) we used Lasso as a screening technique. From there we used the variables for the min lambda and then we used the Bic criterion to choose our variables. Finally, because of high multicollinearity and because some variables were not statistically significant some of them were dropped. In this model we did not use the state variable. The Pseudo R squares had poor performance, meaning that the model did not fit well the data.

### 4.1.2 Second Model

In the second model (Table 3, in the appendix) we did not used the state variables, we logged those that had values more than one thousand and later we used the Bic criterion. Finally, we removed all variables that cause multicollinearity and those that were statistically unsignificant. The Pseudo R squares had poor performance, meaning that the model did not fit well the data.

### 4.1.3 Third Model

In the third model (Table 4, in the appendix) we decided not to use any criterion, since we do not have many variables in the first place. We removed those variables that had high correlation between them and in the end, we removed the statistically unsignificant variables. The Pseudo R squares had poor performance, meaning that the model did not fit well the data.

### 4.1.4 Fourth model

In the fourth model (Table 5, in the appendix) we first removed the predictors with high linear correlation. Then we used the BIC criterion for the variable selection and finally we removed any variables that were statistically unsignificant. In this model the Pseudo R squares and the Hosmer-Lemeshow Goodness-of-Fit Test had better results than the previous three models.

### 4.1.5 Fifth model

In the fith model (Table 6, in the appendix) we used the exact same steps as the fourth model but this time we used the log for those predictors that had values larger than one thousand. The model had the approximately the same Pseudo R squares and Hosmer results.

### 4.1.6 Sixth model

In the sixth model (Table 7 in the Appendix) we used the stepwise with BIC criterion from the null to the full model. The result are promising. The Pseudo R squares have good results, and the Hosmer test has even better results than the fourth and fifth model. We can see a big improvement in the models when incorporating the state column as a predictor

**Table 9: Models comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **CoxSnell** | **Nagelkerke** | **McFaddenAdj** | **P-value**  **Hosmer-Lemeshow Test** | **Chi-squared**  **Hosmer-Lemeshow Test** |
| Model 1 | 0,13 | 0,18 | 0,10 | 0.143 | 12.180 |
| Model 2 | 0,14 | 0,19 | 0,11 | 0.460 | 7.730 |
| Model 3 | 0,16 | 0,22 | 0,12 | 0.942 | 2.871 |
| Model 4 | 0,57 | 0,78 | 0,62 | 0.166 | 11.679 |
| Model 5 | 0,57 | 0,78 | 0,62 | 0.177 | 11.468 |
| Model 6 | 0,57 | 0,78 | 0,62 | 0.308 | 9.419 |

A good fit model should have high Pseudo R squares, high p-values and small Chi-squared, for the Leeshawn Test. From the model comparison we can conclude that the appropriate model is the model 6. With that in mind model 6 will be the Main model for our analysis. For the model 6, the Homser-Lemeshow Test has a p-value > 0.05, we do not reject the null hypothesis, therefore the model fits the data well.

## 4.2 The Main Model

### 4.2.1 The Model

Model 6 was chosen as the main model for this paper. We had issues with multicollinearity which was solved. Also, the BIC criterion was used to choose the predictors for the model. From Table 9 we can see that the result are the best in comparison to the other models. We have a good fit and the model can be used for inference. Regarding the Model 6 the glm function, vif scores and rest of R code can be found in Table 8 in the appendix.

**Model:**

Logit(trump\_dominance) = -1.438e-01\* AGE295214 + 1.663e-04\* INC910213 + 1.581e-01\* PVY020213 + 7.913e-02\* AGE775214 - 2.235e-01\* EDU685213 + 5.713e+00\* state\_id29 -1.485e+00\* state\_id26 + 4.286e + 00\* state\_id24 + 8.392e+00\* 8.392e+00 - 2.012e+00\* state\_id16 + 5.899e+0\* state\_id15 - 1.810e+00\* state\_id14 - 1.804e+00\* state\_id13 + 3.574e+00\* state\_id11 - 1.511e+00 \* state\_id10 - 3.176e+00\* state\_id3

Null deviance: 3577.0 on 2710 degrees of freedom

Residual deviance: 1250.5 on 2668 degrees of freedom

AIC: 1336.5

**Table 10: State id to State Name**

|  |  |
| --- | --- |
| **State ID** | **State Name** |
| state\_id3 | Arkansas |
| state\_id10 | Illinois |
| state\_id11 | Indiana |
| state\_id13 | Kentucky |
| state\_id14 | Louisiana |
| state\_id15 | Maryland |
| state\_id16 | Michigan |
| state\_id20 | Nebraska |
| state\_id24 | New York |
| state\_id26 | Ohio |
| state\_id29 | Pennsylvania |

### 4.2.2 Analysis and coefficient interpretation

**Interpretation of continuous variables:**

* 1 unit increase in the persons under 18 years old, as percentage, decreases the log odds of trump dominance by 0.1438, assuming that all other variables are constant.
* 1 unit increase in per capita money income increases the log odds of trump dominance by 1.663, assuming that all other variables are constant.
* 1 unit increase in persons below poverty level, as percentage increases the log odds of trump dominance by 0.1581, assuming that all other variables are constant.
* 1 unit increase in persons 65 years and over, as percentage, increases the log odds of trump dominance by 0.07913, assuming that all other variables are constant.
* 1 unit increase in the percentage of persons age 25+ with bachelor’s degree or higher, decreases the log odds of trump dominance by 0.2235, assuming that all other variables are constant.

**Interpretation of categorical variables:**

* The log odds of trump dominance is reduced by 3.176 if the voting state is Arkansas.
* The log odds of trump dominance is reduced by 1.511 if the voting state is Illinois.
* The log odds of trump dominance is increased by 3.574 if the voting state is Indiana.
* The log odds of trump dominance is reduced by 1.804 if the voting state is Kentucky.
* The log odds of trump dominance is reduced by 1.810 if the voting state is Louisiana.
* The log odds of trump dominance is increased by 5.899 if the voting state is Maryland.
* The log odds of trump dominance is reduced by 2.012 if the voting state is Michigan.
* The log odds of trump dominance is increased by 8.392 if the voting state is Nebraska.
* The log odds of trump dominance is increased by 4.286 if the voting state is New York.
* The log odds of trump dominance is reduced by 1.485 if the voting state is Ohio.
* The log odds of trump dominance is increased by 5.713 if the voting state is Pennsylvania.

## 4.3 Assumption Control

The assumption that is to be checked is the **linearity assumption** of the logit. To test this assumption for the main model, we run the logistic regression again but this time we include the log interaction of each predictor. What we are looking for is statistically significant on the log term of each predictor. If it is statistically significant, we can conclude that the assumption of linearity is not met (Table 11 in the Appendix).

From table 11 in the Appendix, we can conclude that the assumption of linearity is not met for the variables AGE775214 and AGE295214.

The second assumption to be checked is the **multicollinearity assumption**. In our case the ‘’ GVIF^(1/(2\*Df)) ‘’ for all variables is less than 3.16 which from some statisticians is a threshold. Also, if we run the linear regression as seen from table 12 in the appendix the VIFs scores are small, less than 10, indicating the lack of multicollinearity.

We will not discuss about the **outliers** since this research is beyond this paper analysis and it will take more time to come in strong statistically conclusions.

# Conclusions

The three starting models, those that do not incorporate the state id as categorical variable show small Pseudo R squares. With the incorporation of the state id we see a big increase in Pseudo R squares, meaning that the states help explain the depend variable. Most of the logistic regression assumptions is met but we have some departure from the linearity assumption. We found five variables and eleven states statistically significant. Also, the intercept was found statistically unsignificant.

For future statistical analysis, the outliers should be delt with and also the use polynomials may help the linearity assumption to be fully met. But, we have to take into account that with the use of polynomials the interpretation will be significantly different and harder to explain.

# Appendix

**Graph 3: Correlation Plot**

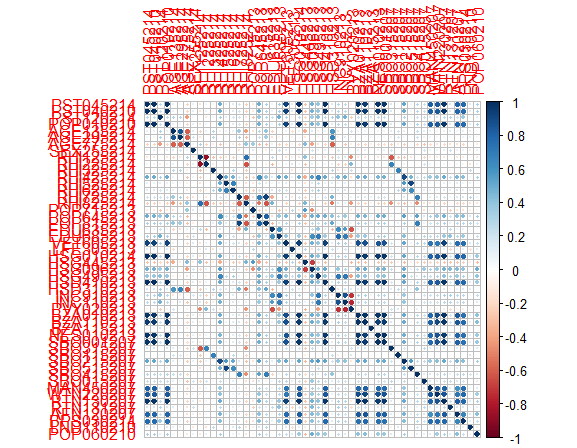


Table 2 first model:

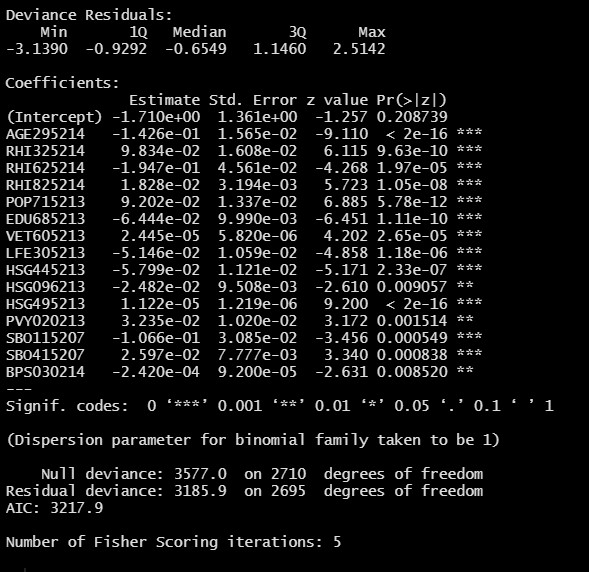


Table 3 second model:

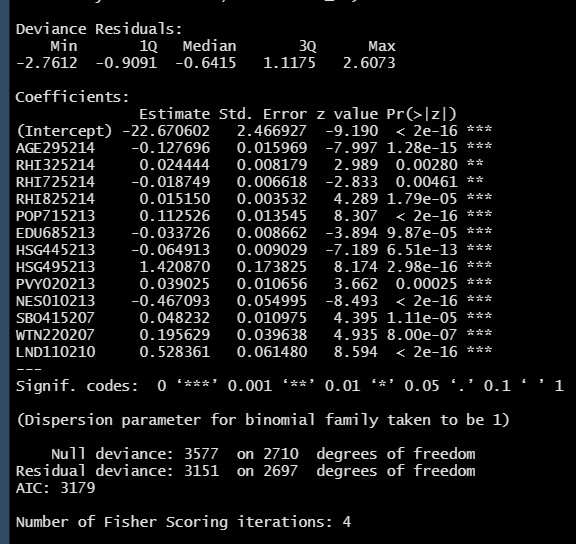


Table 4 third model:

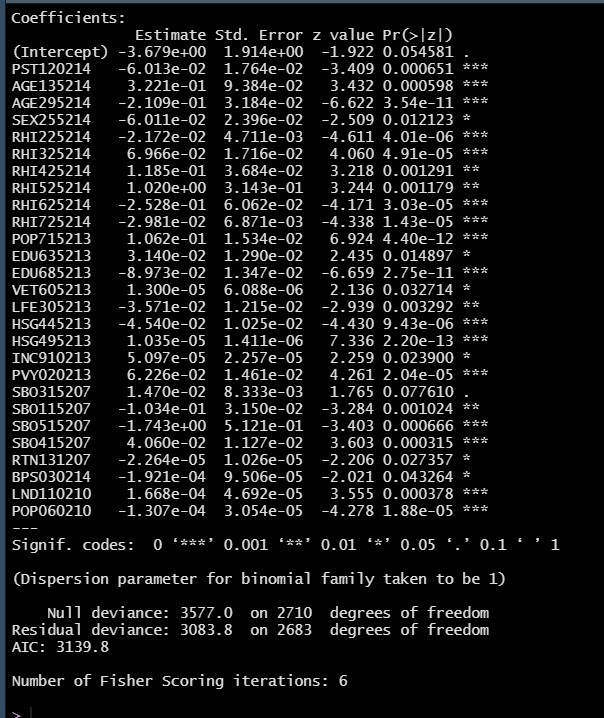


Table 5 fourth model:

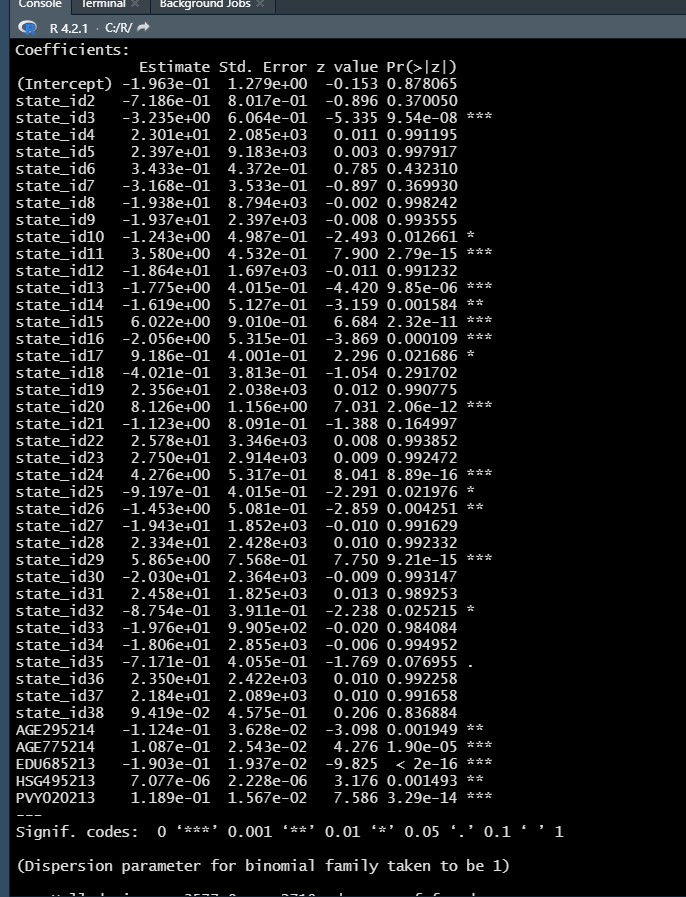


Table 6 fifth model:

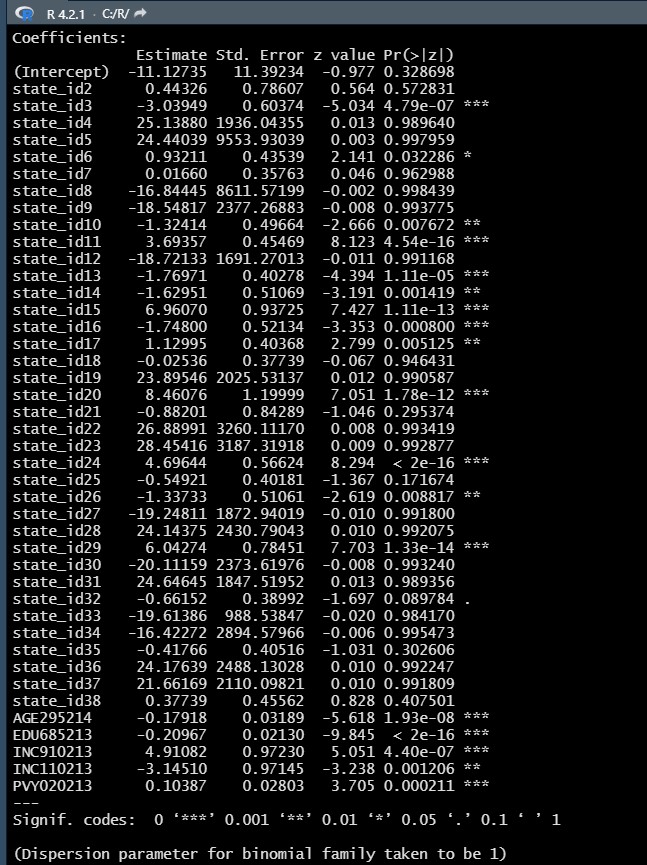


Table 7 sixth model:

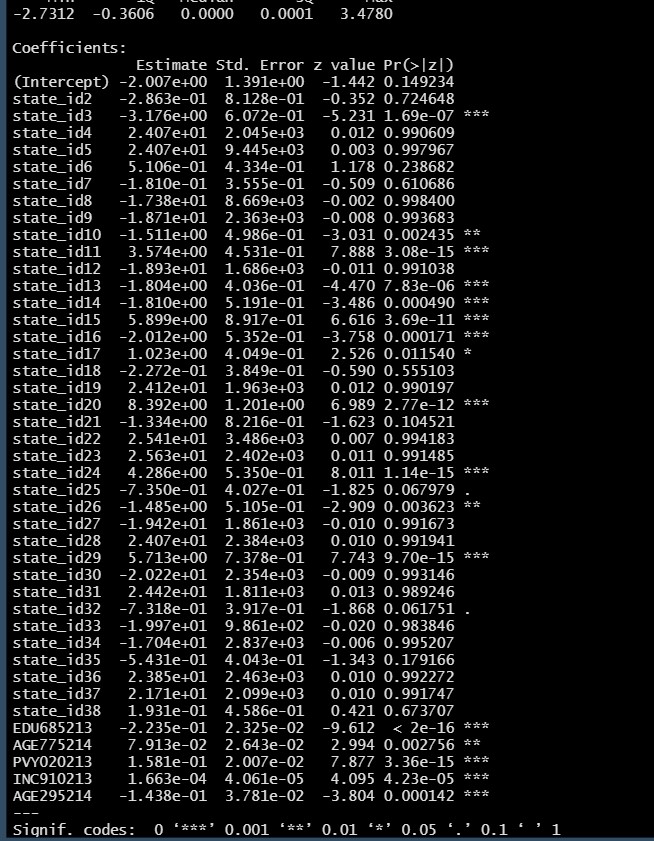
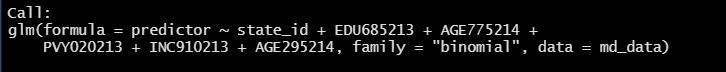


Table 8, model 6: vif, Pseudo R squares, Hosmer-Lemeshow Test and glm function:



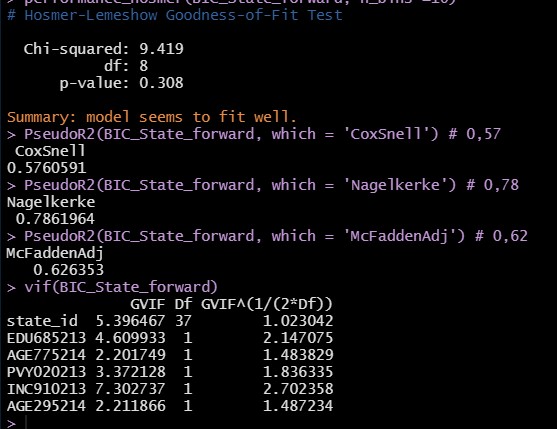


Table 11: interaction terms in main model:

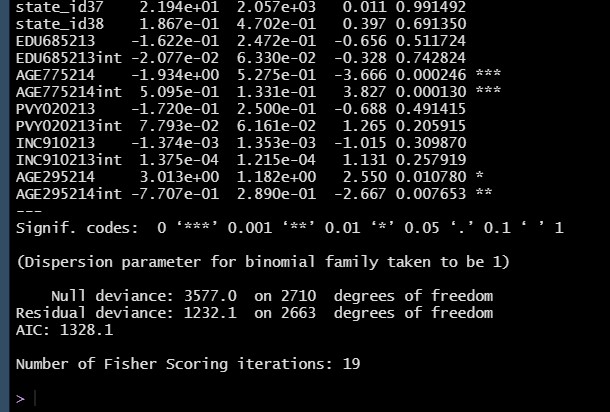


Table 12: Checking the VIFs for the continuous variables of the main model

